

DOCUMENT RESUME

ED 120 290

UD 015 819

AUTHOR Ebel, Robert L.
TITLE Constructing Unbiased Achievement Tests.
PUB DATE Dec 75
NOTE 25p.; Paper presented at the National Institute of Education Test Bias Conference (Annapolis, Maryland, December 2-4, 1975)

EDRS PRICE MF-\$0.83 HC-\$1.67 Plus Postage
DESCRIPTORS *Achievement Tests; Criterion Referenced Tests; *Culture Free Tests; Educational Testing; Language Variation; Measurement Instruments; Non English Speaking; Reliability; Research Problems; Standardized Tests; *Test Bias; *Test Construction; *Testing Problems; Test Interpretation; Test Results; Test Selection; Test Validity

ABSTRACT

The approach characterizing this study of bias in tests of educational achievement is stated to be reflective and philosophical rather than experimental and statistical. Its concern is with operationally definable areas of learning and indicators of achievement rather than with the hypothetical latent traits that are presumed by some to influence or determine test performance. A brief discussion of emotionality in test bias, that also considers the seriousness of the problem, is followed by comments on bias in selection testing. The meaning of bias in achievement tests is discussed next together with other topics such as: bias in the test itself, the problem of reading difficulty and linguistic difference as a possible source of bias, non-standard English speakers, language specificity in achievement, test bias and test validity, constructing unbiased achievement tests, operational definition versus construct validation, appearance versus reality in test validation, the correctness of operational definitions, and a discussion of two propositions advanced to support the suggestion that criterion referenced tests minimize bias. Issues discussed throughout the paper are noted to suggest that bias in tests of educational achievement is quite improbable. The a priori assumption that an achievement test is essentially unbiased is considered more reasonable than the assumption that it is biased. (Author/AM)

* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *

5
for Eric Clearinghouse
Jean Barabas

ED120290

Constructing Unbiased Achievement Tests*

Robert L. Ebel
Michigan State University

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

Introduction

First, a few general comments on the focus of our discussions in this meeting, and of my approach to it. The matter of bias in achievement tests deserves careful study and thoughtful discussion. No doubt all of us have been reading and thinking about it for a number of years, and have intensified our efforts since this meeting was announced. No doubt we have all benefited from the ideas of others.

While it is difficult for us to acknowledge fully all the help we have received from other educators, I would like to mention three to whom I owe a special debt. Ronald Flaugher¹ has written a perceptive analysis of several popular misconceptions that have contributed to excessive concern over test bias. Ross Green² and his colleagues have produced some sophisticated statistical analyses of data on test bias. My perceptions of the nature of the problem, and of the most appropriate approaches to solving it differ somewhat from Dr. Green's. Nevertheless I have found his insights most helpful. They stimulated a number of new, and I hope useful, reflections on bias in achievement tests. Finally, my colleague John Hunter gave me the benefit of a number of perceptive comments that helped in clarifying or correcting statements made in the first draft of this paper.

What I have to offer are opinions and might even deserve to be called personal opinions were it not for the fact that the suggestions

* Prepared for presentation at the NIE's test bias conference held in Annapolis, Md., December 2-4, 1975.

UD 015819

and data of so many others have contributed to them. They are sound opinions, in my own eyes at least. But they are not universally shared. I offer them, not as truth with a capital T, but as carefully considered conclusions, worthy of your consideration.

An Approach to the Study of Bias

My approach to the study of bias in achievement tests has been more reflective and philosophical than experimental and statistical. This means that the paper will be more concerned with defining concepts clearly and with making rational inferences from reasonable premises than with collecting and analyzing data. It will be more concerned with operationally definable areas of learning and indicators of achievement than with the hypothetical latent traits that are presumed by some to influence or determine test performance. The primary goals will be rationality of thought and clarity of expression. Concreteness is preferred to abstractness, simplicity to complexity, specificity to generality.

Three things can be said in defense of this approach. One is that the first step in dealing with any problem is to define it clearly and analyze its facets. No substantial consensus has been reached on this definition and analysis of achievement test bias. Another is that the conception of achievement in learning, implicit in factor analysis and other latent trait models, involves what seem to me to be questionable assumptions and unverified hypotheses. A third, related to the second, is that the results of elaborate statistical analyses of rather soft data tend to be of questionable reproducibility and generalizability. What one tends to get from those analyses are hints and suggestions rather than solid conclusions.

Whether the philosophical approach to the problem of test bias is better than the statistical, at least it is different. Whether it is what you are prepared to accept, it is about all I have to offer. I am not forgetting that my assignment is to talk about test construction--the construction of unbiased tests. But to defeat the enemy--bias--we need first to locate him and to discover how he operates.

Emotion in Discussions of Test Bias

Discussions of test bias are often tinged, and sometimes even saturated, with emotion. Racial injustice, and the dubious value of some attempts to correct it, can not help but arouse strong feelings. This is both good and bad. It is good in that it motivates and energizes efforts to correct injustice. It is bad because emotion often interferes with or substitutes for clear thinking. Reason sometimes stumbles over emotion. If, in our discussion here, we can keep reason in control and emotion in check we will be likely to reach sounder conclusions.

How Serious is the Problem?

Sometimes the problems that vex and perplex us become much less formidable and troubling when they are examined closely and defined clearly. They may even cease to appear as problems that require urgently to be dealt with. I strongly suspect that this may be the case with bias in achievement tests. If we define what a test is measuring in terms of the operations used to measure it, then a basis for belief that the test is biased is almost impossible to find. What a test of educational achievement asks pupils to show what they can do does not depend on whether they are white or black, male or female, inner city or suburban.

How well they can do it may depend on these and other factors. But this does not keep the test from measuring what they can do.

Denial of the seriousness of the problem of bias in educational achievement tests is not currently a popular response to criticism of tests, even among test authors and publishers. They are much more inclined to say, in essence, "Yes, we know that bias exists in our tests, and we are taking vigorous action to eliminate it." No doubt this kind of a statement reflects the honest conviction of those who have expressed it. No doubt it has stimulated beneficial soul-searching and productive research. But it also serves to protect and defend the public image of test makers and users. It indicates that they have heard the protests of the minorities and other critics of educational tests, and are at work conscientiously to correct the faults. Good intentions are admirable. But no matter how well intended the efforts to eliminate bias from achievement tests may be, they can not succeed if no substantial bias exists to be removed.

Bias in Selection

Note well that this discussion is focused on tests of educational achievement, not on aptitude or ability tests used for selection or placement. The possibilities of bias in selection testing do exist. Biased selection has clearly occurred in a number of instances, sometimes inadvertantly, sometimes intentionally. During the depression years in the 1930's school officials in New York City tested the "intelligence" of applicants for teaching positions by asking them to complete the number series: 14, 34, 42, __, __. This item was intentionally biased in

favor of local applicants. Only they were likely to recognize the numbers as streets at which stops were made by express trains on the Seventh Avenue subway, and that the next two numbers should be 59 and 72.

It is not the function of this paper to discuss bias in selection testing. But in view of the great interest in this aspect of test bias, and the various suggestions that have been made for coping with it, two brief comments may be appropriate here. The first is that, contrary to popular belief and the statements of some minority spokesmen, test bias in academic selection is more often helpful than harmful to minority applicants. That is, aptitude tests tend to predict somewhat greater scholastic success than minority students achieve, on the average.^{3,4,5,6}

The second is that selection strategies such as those proposed by Nancy Cole and Robert Thorndike, intended to reduce cultural bias and improve fairness in selection have two limitations:

- 1) They purchase "fairness" at the expense of "effectiveness" in selection, and
- 2) They should be applied, in the interests of fairness, not only to members of a cultural minority but to anyone, regardless of race or sex, who makes a low score on the test.

The Meaning of Bias in Achievement Tests

With this brief nod in the direction of bias in selection testing, let us return to the question of bias in tests used to measure achievement. What do we mean when we speak of bias in an achievement test? Surely not just that some pupils make lower scores than others on the test. Surely not that the test was badly administered, or the scores misinterpreted, or the pupils unaccustomed to test taking or unwilling to cooperate in being tested. These are not the fault of the test. They can not be corrected by constructing an unbiased test.

An achievement test is biased if, when correctly administered and taken, it results in scores for some of the takers that rank them systematically lower or higher than they ought to be ranked, on the basis of true achievement, among the other test takers. It is bias in the test itself that concerns us here, not in the giving or taking of the test. However, to avoid confusion on this point it may be advisable to note specifically some of the ways in which bias can be introduced in test administration and in test taking.

Test scores may be biased if the test administrator:

- 1) gives some pupils less time than they need to finish the test.
- 2) gives some pupils more help in understanding the directions, or the question, or in deciding on the answer.
- 3) allows some pupils to help each other or to use unauthorized notes.
- 4) frightens some pupils more than others.
- 5) allows some pupils to be distracted or interrupted during the test.
- 6) gives directions that some pupils can not hear, or understand.

Test scores may also be biased if some pupils:

- 1) are more familiar than others with the form of item used in the test.
- 2) are more familiar than others with the method of recording responses.
- 3) distribute the available time over the various tasks in the test more wisely avoiding waste of time on difficult items which might keep them from having time to consider all items.
- 4) make wiser use of opportunities to guess in relation to the procedures to be employed in scoring the tests.
- 5) try harder than others to do well on the test.

To minimize the sources of bias in the giving of tests well trained and conscientious test administrators are essential. To minimize bias in test taking, pupils who are reasonably test wise and reasonably well

motivated are essential.^{7,8} If a pupil is poorly motivated or for some reason does not choose to have his educational achievements measured, no test or tester can measure them. His score will be biased, but it is not a kind of bias that can be blamed justly on the test or the tester. The pupil has mainly himself, or perhaps his cultural background to blame for not allowing the test to measure his achievements as accurately as it measures the achievements of other pupils.

Bias in the Test Itself

What of bias in the test itself? Let us consider three alleged sources of such bias.

- 1) The test requires abilities that not all of the pupils have had an equal opportunity to learn.
- 2) The test makes use of language that some pupils are better acquainted with than others.
- 3) The test actually measures different aspects of achievement in different pupils.

Tests intended to measure basic general ability to learn (I.Q.), often base that measurement on how much has been learned. For such tests it is crucial that all those to be measured have had equal opportunity to learn the knowledge or skills that will be tested. This ideal, incidentally, is never actually realized in practice. It was the obvious and substantial lack of that kind of equality between children of disadvantaged minorities and of the white majority that provided the basis for Allison Davis's⁹ criticism of intelligence tests a quarter of a century ago.

But in testing achievement, that is in attempting to measure how much has been learned of a particular body of knowledge, variations in opportunity to learn it are quite irrelevant. Differences among pupils in opportunity to learn surely exist, and surely affect how much has been learned. But those differences introduce no bias in the measurement of how much actually has been learned.

Children of the inner city no doubt learn things that children of the suburbs do not learn, and vice versa. If pupils from either type of neighborhood do poorly on a particular test of achievement because they had less opportunity to learn what that test requires, one should not conclude that the test is biased. Nor should one conclude that the pupil is generally inferior, mentally or culturally, to other pupils. What can be concluded, if the test is valid, is only that the pupil has achieved less in this particular area of learning.

The Problem of Language

Consider next the problem of reading difficulty and linguistic difference as a possible source of bias in achievement tests. Many pupils do have difficulty in reading. In some cases their oral speech differs quite sharply from standard English. These are serious educational problems, but their seriousness as sources of bias in achievement test scores is easy to overestimate.

To be biased an achievement test score must consistently understate or overstate a pupil's true achievement. If a pupil has difficulty in reading and understanding the test questions it is highly probable that he also had difficulty in reading and understanding the books and other

9

written materials used in instruction. It is most unlikely that he possesses a good store of the knowledge the school tried to teach, but which his inability to read the test questions prevents him from showing.

Linguistic limitations are not merely handicaps to communication. They limit learning itself. For words are much more than the means for communicating ideas. They are the materials out of which ideas are made. Most of the knowledge that fills our textbooks and encyclopedias, our libraries and our minds, is verbal knowledge. Most of our thinking, our reasoning, and our creative imagining is done with words. Reading is not merely a special skill, useful in the process of communication. It can not be separated in any meaningful way from learning, from knowing and from thinking.^{10,11}

But what of the blind? They do not read, at least in the usual way. But they do learn. Is not listening a good substitute for learning? For some it may be almost the only substitute, and a reasonably good one. But a sighted pupil who must rely on listening as the source of information for learning misses a great deal. Unless he makes the very special efforts to learn that some blind persons do, he is likely to learn far less than readers learn.

Non-standard English

Some children come from homes and neighborhoods where standard English is not spoken. If these children are asked to take a test whose questions are expressed in standard English, are they not at a disadvantage? Are not the tests biased against them? "Indeed yes!" say some critics of achievement tests. But is it reasonable to suppose

that the linguistic barrier that hampers their test taking has not also hampered their school learning? Is it reasonable to suppose that if the test were to be translated somehow into a written expression of the spoken non-standard English with which the child has become familiar that he could read the questions with understanding? Is it reasonable then to suppose that he could answer the questions correctly, drawing on an adequate structure of knowledge he has built out of non-standard terms, and of propositions expressed in non-standard syntax? Both common sense and research^{12,13} indicate that the answers to these questions are all negative.

So far as I know, there is no substantial body of human knowledge that has been written in non-standard English. Beyond doubt it is helpful for the teacher of children who have learned to speak non-standard English to be able to understand what they are saying, and to speak to them in a language they understand. But if the children do not quickly learn to communicate in standard English; to read and write it, speak it, and understand it, their future years in school and adults will surely be difficult, discouraging and unproductive. The business of not only the student, but of the worker and the citizen is conducted mainly in standard English.¹⁴

There is no empirical evidence I know of to support the belief that achievement tests in standard English yield scores that underestimate the actual achievements in school learning of children who speak and understand non-standard English. There is no good way I can think of to get such evidence. Lacking evidence one must rely on rational inference. The considerations set forth above suggest that the bias, if it exists

at all, is probably very small. They suggest this also. The pupil who scores low on an achievement test because of his unfamiliarity with standard English reveals a serious deficiency in achievement that urgently requires correction. His low score truly signals low achievement. It is not a biased measure of achievement.

Language Specificity in Achievement

Most educational achievements are language-specific to a very considerable degree. An expert accountant who speaks only English can not use his expertness to serve a business whose managers speak only Japanese. An expert chemistry teacher who speaks only German can not transmit his expertness to students who understand only English. It is not just that language is a means of communication, which surely it is. The language is a large part of the substance of the knowledge itself. Verbal knowledge, which is what almost all of our school-acquired knowledge is, exists initially only in the language in which it was learned. If a person manages to translate it, at the cost of great labor, from the language in which he learned it into another in which he hopes to use it, he comes to possess two bodies of knowledge, related but separate. When he thinks about that knowledge, it will be in one language or the other, never in the abstract.

Consider a test of knowledge of the solar system intended for twelve year old pupils and written in English. Is it biased against a French speaking boy of twelve years who studied the solar system in the schools of Lyon? It might be said to be so, since his score would be a poor indication of how much he knew of the subject. Yet if bias is taken

to imply a correctible fault in the test, this test is not biased since a test free of this kind of fault is clearly impossible. The test must be written in some one language. Would it not be much more accurate to designate the test as a measure of knowledge in English of the solar system? A low score on such a test is not a biased measure of the French boy's knowledge in English of the solar system.

Test Bias and Test Validity

This brings us close to a topic that needs to be considered in any discussion of test bias; the topic of validity as it relates to achievement tests. L.L. Thurstone¹⁵ spoke for many other test specialists when he said "By validity is meant the correlation between a test and its criterion." For some of these specialists no other conception or definition of validity is necessary. But what of tests of educational achievement? Where or what is the criterion for judging the validity of such tests?

One may say that it is what the test was intended to measure. But if what the test was intended to measure has not been operationally defined, such as criterion has no real existence or practical usefulness. It yields no quantitative measures of what the test was intended to measure. No correlation coefficient can be calculated from it. It provides no measure of test validity. Thurstone's definition of validity, it would appear, has no operational relevance to educational achievement tests.

This might lead one to conclude that for such tests the question of validity need never arise. But there is another, perhaps better conclusion.

It is that test of educational achievement requires a somewhat different conception and definition of validity. Here is one possibility.

A test of educational achievement is valid to the extent that it:

- 1) is composed of tasks which sample representatively some clearly defined domain of tasks, and
- 2) yields reliable scores.

The "defined domain of tasks" indicates as clearly as can be indicated what the test is intended to measure. The representativeness of sampling of the tasks included in the test, and the reliability coefficient of the scores indicate how well the test measures what it was intended to measure.

If a test is valid in this sense it can not possibly be a biased test no matter who takes it. A physician's weight scale is not biased for or against children, or women, or Blacks. A test of ability to spell correctly from dictation of English words taught in a particular fifth grade is not biased against young geniuses, or slow learners, or Chicanos, or refugees, or late transfers from another school who were not taught most of those words. It measures what it was intended to measure.

Constructing Unbiased Achievement Tests

How then does one construct an unbiased test of educational achievement.

- 1) Clearly define a domain of tasks.
- 2) Give the test a name which indicates accurately the principal features of that domain, and differentiates that test clearly from other tests of the same general kind.

- 3) Sample representatively from that domain.
- 4) Calculate the reliability of scores on the test from a representative sample of the population of intended examinees.

If the first three requirements are reasonably well met, the test is valid to the extent that the scores are reliable. If it is reasonably valid in this sense, it is reasonably unbiased.

Operational Definition vs. Construct Validation

Our position in this paper is that the question of validity in achievement tests, and the related question of bias, are matters that can best be dealt with by indicating clearly what the test is measuring. The characteristic the test is intended to measure is operationally defined by the process of test development--the nature of test tasks, the content areas, the language context, the procedures for administration and scoring. What is being measured is defined in terms of evident behavior in response to test tasks, not in terms of latent traits which presumably determine that behavior.

There is another approach to the validation of achievement tests. Instead of basing the claim for validity on an operational definition of the thing to be measured, the claim is based on the correlations found between scores on the test and scores on other tests or measurement procedures. From a study of these correlations, it is said, the test maker can determine whether his test actually is measuring what he hoped to measure. This procedure is referred to, somewhat loosely I think, as construct validation. It has little in common with the kind of construct validation Cronbach and Meehl¹⁶ had in mind when they introduced and

developed the concept. These test centered constructs tend to be poorly defined observable abilities rather than hypothetical constructs. They are not parts of any coherent theoretical structure. Seldom are they involved in exact hypothetical relationships with other variable that can be tested experimentally. I know of no test of educational achievement whose essential nature, whose freedom from or contamination by bias, has been or needs to be elucidated by construct validation procedures.

Direct test validation based on an operational definition of what the test is intended to measure does not satisfy all test users. Where is the evidence of its validity? These persons feel much more comfortable with indirect validation based on criterion measures. But the criterion measures must also be valid, and usually the kind of validity they have is direct validity based on some implicit or explicit operational definition. Thus, there is no escape ultimately from direct validation of either test or criterion. The evidence of direct validity is provided by the test itself, or by the source of the criterion measures. When, on inspection, the tasks in the test appear to require the ability the test is intended to measure, the test is judged to be valid. If this seems a shaky and uncertain basis for claims of validity it may help to recognize that this is the only available ultimate basis for validating any process of measurement. How else were measures of typing speed, temperature, molecular weights, consumer price indexes, height above sea level, or inter-stellar distances validated? Even construct validity, so called, requires operationally defined variables to produce measurements which can be correlated.

Appearance vs. Reality in Test Validation

But, says the persistent skeptic, how do you know that the tasks in your test really require the ability you intend to measure? How do you know on what basis a pupil chose the correct answer, or what caused him to miss it? Beyond that, how do you know that the operational definition you have formulated or accepted is the correct definition?

The possibility that a test item may not function as it was intended to function can not be denied. Clues provided by grammatical structure (e.g. use of both singular and plural forms in responses to a multiple-choice stem which calls for a singular response) by past verbal associations (e.g. use of "Interstate Commerce Commission as a response to a multiple-choice stem asking which government agency regulates commerce between the states), or by overly specific questions (e.g. asking which Frenchman explored Lake Michigan in an item in which only one of the men listed has a French name) can enable one who lacks the knowledge being tested to respond correctly. On the other hand lack of specificity in the question or ambiguities in the phrasing can prevent one who has the knowledge from responding correctly.

One clear example of test items that do not function as they were intended to function is provided by certain reading test items. These items are intended to test a student's comprehension of a reading passage. Some of them, however, can be answered as well by students who do not have the passage to read at all. Such items are not "passage dependent."¹⁷ They may test comprehension of the question, but not of the reading passage given to be comprehended.

Usually the failure of an item to function as it was intended to function is a result of faults in item writing that a critical eye can catch, or that a study of responses to the item can suggest. They can be corrected most directly and effectively by revising the items, not by attempting to determine through correlational analyses what the item is "really" measuring. To write test items that do in fact measure what they were intended to measure and what they appear to measure is not an inordinately difficult, and certainly not an impossible task.

The "Correctness" of Operational Definitions

The problem of "correctness" of the operational definition can be dealt with by recognizing it as a pseudo-problem; one for which no solution can ever be found. For the problem arises from the false assumption that the true meaning of a human trait or characteristic exists somewhere to be discovered. This is no more true of traits than it is of words or of political boundaries. What we need is not an operational definition that is "correct" but one that is generally acceptable. For until it has been operationally defined an achievement can not be expressed as a quantitatively meaningful variable.

We sometimes speak of human achievements such as reading ability, spelling ability or reasoning ability as if they exist as meaningful quantities quite apart from our attempts to measure them. A good speller, for example, is expected to do well on any kind of a spelling test--dictation, multiple-choice, or error-avoidance in free writing. If a person does better on one kind of test than on another, the difference is attributed to imperfections, that is invalidities, in one or more of

the tests as measures of a person's "true" spelling ability. But the tasks in these three kinds of tests are somewhat different. One requires production of prescribed words, correctly spelled. Another requires recognition of correctly and incorrectly spelled words. A third requires sensitivity to possibly misspelled words, and avoidance of them or verification of the correct spelling. While one would expect a fairly high correlation between scores on these three tests, one would not expect it to be perfect, even if the scores would be made perfectly reliable. For the three are not tests of precisely the same ability to perform. They measure somewhat different abilities. What each measures is defined operationally by the tasks it presents.

The case of reading is similar to that of spelling. There is not just one reading ability, there are many. There is the ability to read stories for young children, ability to read a research report on lymphocyte surface immunoglobulins, ability to read the dialogues of Plato or the poems of Gertrude Stein. In more general terms there is oral reading, reading for information storage,¹⁸ and reading for comprehension. Somewhat different operational definitions of reading are involved in measurements of the last three, and possibly of the first four as well. It is impossible to argue reasonably that any one of these operational definitions is the correct definition of reading. Each is meaningful in its own way, and useful for some particular purposes.

Do Criterion Referenced Tests Minimize Bias?

The suggestion has been made that criterion referenced achievement tests (CRT) might be less subject to bias than norm-referenced tests (NRT)

presumably are.¹⁹ Two propositions are advanced to support this suggestion.

1) Scores on criterion referenced tests depend more on specific learnings and less on general background knowledge. 2) Differences in general background knowledge are responsible for bias in achievement tests. The basis for both of these propositions is rational inference, not empirical evidence. Considered rationally, the second proposition seems to have somewhat greater credibility than the first.

The more an achievement test score is influenced by the examinee's general knowledge, the less valid it is likely to be as a measure of some specific achievement in learning. Such a test would be clearly biased against any student of low general knowledge but high specific achievement. Students of this kind are not likely to be very numerous, however. Most students tend to be consistently better or poorer at specific learnings, and every pupil's general background is a result of a very large number of specific learnings. So while the second proposition is credible on theoretical grounds, it is unlikely to be the cause of much actual bias.

The first proposition is less logically persuasive. Green has argued that even though the items for a criterion-referenced tests are indistinguishable individually from those for a norm-referenced test there are two characteristics of the criterion-referenced test as a whole that make it less sensitive to general background. 1) The items in a CRT do not deal with a broad range of objectives. 2) The score on a CRT is obtained by counting objectives reached, not items answered correctly. True, these are characteristics of some, if not most CRT's. But how do they protect CRT scores from the influence of general background? The

area of achievement to be measured by a NRT can be defined clearly enough to enable the test constructor to identify inappropriate items or item characteristics. The range of objectives in a CRT of achievement in any school subject must be as broad as that for a NRT for the same subject. Otherwise the two are not tests of the same achievement. And how does counting objectives instead of items exclude or minimize the influence of general background?

Because criterion referenced tests are new, and unfamiliar, and somewhat different; because they are not just ordinary tests, but involve the latest advances in the technology of testing, they can be, and have been promoted as a panacea for most of the alleged shortcomings and evils of ordinary tests; for indefinitives in the content covered by the test, for deficiency of meaning in the test scores, for invidious comparisons and destructive competition among pupils, for failures to learn and discouragement of efforts to learn, and for bias in the tests. There is some truth in some of these claims, but much less than the promoters have implied. And CRT's have their own problems and limitations, in some cases rather severe. CRT's are very popular with those who are looking for a painless alternative to testing. But CRT's, highly useful in particular situations especially in the elementary grades, will never provide a satisfactory solution to most of our testing problems. It is my considered opinion that neither norm-referenced nor criterion referenced tests of achievement are afflicted with appreciable bias. If NRT's were to be so afflicted, I see no reason to believe that CRT's would be less biased.

An Operational Definition of Test Bias

In concluding this discussion of bias in achievement tests it may be appropriate to note that no direct evidence of the kind of achievement test bias defined in this paper seems to have been reported. All of the concerns that have been expressed seem to be based on suspicion, rumor, or hypothetical assumption. It is indubitably true that members of some groups make lower scores on a particular achievement test than do members of other groups. But surely this is not sufficient evidence of bias in the test. It is possibly true that the same test will yield scores having a somewhat different factor structure in one group than in other groups. But this is even more tenuous evidence of bias in view of the tentativeness and variability of the conclusions reached in studies attempting to determine "factor structure." Has the factor structure of any test of educational achievement been uniquely and finally determined?

How might one obtain clear direct evidence of bias? Earlier in this paper a biased achievement test was defined as one that gives some of the examinees higher or lower scores than they ought to get. Can this conceptual definition be operationalized? If it can, we will be in a position to determine whether, and to what extent, an achievement test is biased.

What is required clearly is some means of knowing what score each examinee ought to get. The only way of getting such scores would be from some unbiased test. To tell whether Test X is biased, and if so to what extent, we need Test A which is unbiased. This begins to sound like a paradox. To get what we want (an unbiased test) we must already have it!

Actually, the problem is not wholly insoluble. Given two different tests of the same achievement which yield systematically different scores for one or more subgroups of the examinee population, the possibility exists that one of the two is biased. Examination of specifications for the two tests, and of the tests themselves, might lead to the conclusion that one is a purer, more valid measure than the other of the desired achievement. The other would then have to be regarded as the more biased.

Alternatively one might correlate scores on the two tests with other examinee characteristics clearly irrelevant to the achievement being measured. If the two tests are equally reliable, the one showing higher correlation with irrelevant achievements is probably the more biased.

If one of the tests is unbiased, or much less biased than the other, one might measure the bias of the other in this manner. Correct the correlation between the two for attenuation. Then express the disattenuated correlation as a coefficient of alienation. This could provide an estimate of the degree of bias in the more biased test.

Obviously direct evidence of bias in an achievement test will not be easy to get. Unless there is solid ground for suspicion that a test may be biased, efforts to verify the suspicion may well be unwarranted. Considerations discussed earlier in this paper led to the conclusion that bias in tests of educational achievement is quite improbable. An a priori assumption that an achievement test is essentially unbiased seems much more reasonable than the assumption that it is biased. Unless and until direct evidence of substantial bias is produced, let us direct our energies not to developing unbiased achievement tests but to developing generally more valid tests.

References

1. Ronald Flaughner, "Some Points of Confusion in Discussions of the Testing of Black Students" The Testing of Black Students (Lamar P. Miller ed.) Prentice-Hall, Inc. 1974.
2. Donald Ross Green, "What Does It Mean to Say a Test is Biased?" AERA Paper, Washington, D.C., March 31, 1975. Unpublished.
3. Julian C. Stanley and Andrew C. Porter, "Correlation of Scholastic Aptitude Test Scores with College Grades for Negroes Versus Whites" Journal of Educational Measurement, Winter, 1967, pp. 119-218.
4. Anthony Kallingal, "The Prediction of Grades for Black and White Students at Michigan State University" Journal of Educational Measurement, Winter, 1971, pp. 263-65.
5. T. Anne Cleary, "Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges" Journal of Educational Measurement, Summer, 1968, pp. 115-24.
6. Michael C. Pfeifer and William E. Sedlacek, "The Validity of Academic Predictors for Black and White Students at a Predominately White University" Journal of Educational Measurement, Winter, 1971, pp. 253-61.
7. Carl Callenbach, "The Effects of Instruction and Practice in Test Taking Techniques Upon the Standardized Reading Test Scores of Selected Second Grade Students" Journal of Educational Measurement, Spring, 1973, pp. 25-30.
8. Michael E. Erickson, "Test Sophistication: An Important Consideration" Journal of Reading, November, 1972, pp. 140-144.
9. Kenneth Eells and Allison Davis, Intelligence and Cultural Differences, University of Chicago Press, 1951, 388 pa.
10. Edward L. Thorndike, "Reading as Reasoning: A Study of Mistakes in Paragraph Reading" Journal of Educational Psychology, June, 1917, pp. 323-332.
11. Robert L. Thorndike, "Reading as Reasoning" Reading Research Quarterly, Vol. 9, No. 2, 9:135-147, 1973-74.
12. Vernon C. Hall and Ralph R. Turner, "The Validity of the 'Different Language' Explanation for Poor Scholastic Performance by Black Students" Review of Educational Research, Winter, 1974, 44:69-81.

13. Walter H. Macginitie, "Testing Reading Achievement in Urban Schools" The Reading Teacher, October, 1973, 27:13-21.
14. William Raspberry, "Should Ghettoese Be Accepted?" Today's Education, April, 1970, 59:30-31.
15. L. L. Thurstone, The Reliability and Validity of Tests, Edwards Brothers, Ann Arbor, 1939.
16. Lee J. Cronbach and Paul E. Meehl, "Construct Validity in Psychological Tests" Psychological Bulletin, 1955, 52, 281-302.
17. J. Jaap Tuinman, "Determining the Passage Dependency of Comprehension Questions in Five Major Tests" Reading Research Quarterly, 9:206-223, No. 2, 1973-74.
18. Ronald P. Carver and Charles A. Darbey, Jr., "Development and Evaluation of a Test of Information Storage During Reading" Journal of Educational Measurement, 8:33-44, Spring, 1971.
19. Donald Ross Green and Glenn E. Roudabush, "An Investigation of Bias in a Criterion-Referenced Reading Test" AERA Paper, Washington, D.C., March 31, 1975.

JUN 24 1976
 11:24